CUSTOMER REVENUE PREDICTION FROM GEOGRAPHICAL DATA

by

Kasra Khademorezaian

July, 2019

Director of Thesis: Nasseh Tabrizi, PhD

Major Department: Computer Science

Online stores have created more opportunities for firms to offer different products and services to their customers. These online stores produce a tremendous amount of data that serve different purposes, including revenue prediction. Online stores usually keep a log of customers that visit their website that include session information, the products they showed interest, IP, location, and device information. These features are explored massively for studies such as customer churn and recommender systems but using location information for prediction is not explored as much.

The first part of this thesis systematically reviews the articles on revenue prediction with respect to their publication date, application area, evaluation criteria, and technique for prediction that provides a good understanding of already conducted research, the evolution of the topic over the years, and possible research opportunities.

The second part focuses on the prediction of Google store revenue data. Using linear regression as a baseline, it evaluates the predictive power of different machine learning techniques, including gradient boosting, support vector regression, and neural networks. The data is collected from Google Analytic demo account that contains 903,653 observation and 55 features. The goal of this study is to predict the total transaction per user from December 1st, 2018 to January 31st, 2019 and in order conduct performance analysis between different prediction techniques.

CUSTOMER REVENUE PREDICTION FROM GEOGRAPHICAL DATA

A Thesis

Presented to The Faculty of the Department of Computer Science

East Carolina University

In Partial Fulfillment of the Requirements for the Degree

Master of Science in Computer Science

by

Kasra Khademorezaian

July, 2019

CUSTOMER REVENUE PREDICTION FROM GEOGRAPHICAL DATA

by

Kasra Khademorezaian

APPROVED BY:

DIRECTOR OF THESIS:

Nasseh Tabrizi, PhD

COMMITTEE MEMBER:

Venkat Gudivada, PhD

COMMITTEE MEMBER:

Haiyong Liu, PhD

CHAIR OF THE DEPARTMENT
OF COMPUTER SCIENCE: Venkat Gudivada, PhD

DEAN OF THE
GRADUATE SCHOOL: Paul J. Gemperline, PhD

Table of Contents

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

There are many reasons to think about revenue prediction. Firms need to know how much revenue they make in order to adjust their marketing, production, or financial strategies. In many cases, these strategies are related directly to individual customers and firms are interested in finding how much an individual client generates revenue or in other words, how much is a customer worth for the firm.

With advances in web services and online stores, firms can collect a variety of data from their customers ranging from their entry link to their devices and their locations. These datasets provide an opportunity to predict not only the total revenue but also the revenue generated by each customer.

Depending on the industry and type of product or service, there are different approaches to revenue prediction. Advances in computation power and ready to use packages facilitate the adoption of sophisticated techniques to obtain a more accurate prediction; however, accuracy is not always the objective where interpretation is often as crucial. The balance between accuracy and interpretability is one of the reasons that feature selection, along with the appropriate modeling technique, is fundamental to a legitimate prediction.

Machine learning techniques have become very popular due to advances in computation power, accessibility to ready to use software packages, and availability of large

datasets that need processing. These models, especially if combined, can achieve high accuracy, handle many features, and discover generalizable patterns. However, the applications of these models in economics are limited due to the fact that models are prediction tools, and many economic applications seek parameter estimation [37].

Revenue prediction is a broad topic that fits into many areas of application, and it is reasonable to perform it using machine learning techniques. These techniques are different based on the nature of the revenue, whether it is movies, goods, services, or tax. Moreover, it is possible to choose the features to utilize in the prediction selectively that allows not only to use machine learning capabilities to acquire a sound prediction but also end up with a reasonably interpretable model.

Selecting features by itself is a broad topic in computer science;however, the focus of this thesis is to select the features that create more interpretable results by investigating the data helps to select a feature set. In this thesis, we used the data set from google store, which is an online hardware retailer operated by Google. Serving customers worldwide, plus the availability of the data raises the question if location information of the customers is a powerful feature set to predict the firm's revenue.

There is a relationship between the customer location and their purchases hence the revenue. Customers from developed countries and those from larger cities often expend more on goods, especially on technology-based goods. The relationship between city size and income [19], land rents, expenditure on public goods [2], and productivity [47] are a recognized phenomenon in urban economics.

This thesis explores and categorizes published literature based on their application area, the modeling technique, and evaluation criteria to find suitable modeling approaches for revenue prediction. These techniques are then compared based on their performance to validate the possibility of revenue prediction from location information.

1.1 Thesis structure

This thesis is structured as follow. Chapter 2 presents a systematic literature survey on revenue prediction. The proposed methodology is described in Chapter 3. Chapter 4 presents the performance analysis of different models using the Google store dataset. Chapter 5 concludes this research and provides future research opportunities.

1.2 Research Contribution:

This thesis conducts a systematic literature review on revenue prediction of published articles and categorizes them base on their application area, prediction techniques, and evaluation criteria. This review also highlights the limitations and drawbacks in this field that need further work.

The second contribution of this thesis is implementation of the geographical data from the logs used as the features to predict revenue. The typical approach to use data with many features is to select the most correlated features, and it is usually done automatically by the algorithm used for prediction. This method of feature selection assures high accuracy, but the results of prediction are not easily interpretable. Increasing the interpretability of a machine learning model was the primary purpose of using geographical data.

The third main contribution of this thesis is the comparison between econometrics models and machine learning techniques. This comparison shows the differences between econometrics and complicated machine learning models that binds the gap between econometrics and computer science.

Chapter 2

Systematic Literature Review

Predicting the generated revenue of the firms is an important topic in many fields, including movies, tax, services, energy, and goods. Motivated by the increasing number of online stores that store a log of all their customers' interactions on their websites, such as their purchase history and geographical location, there exists a need to explore the usage of these abandoned data to revenue prediction. Currently, there is no comprehensive literature survey that encompasses the work already performed in this field, this thesis conducts a systematic review of revenue prediction literature to find some trends and application areas and available techniques that authors used to realize their prediction.

Conducting literature on revenue prediction is not easy as different fields working on predicting revenue use different phrases such as "customer management and marketing", "production management and design", and "economics". This diversity in the fields of study not only makes it difficult to find all articles related to revenue prediction but also shows a necessity to clearly define the revenue before starting a systematic review.

In searching for related literature, there are several phrases and keywords that imply the same meanings, but because of the specific methodology in the study, they fit in revenue prediction. Examples of these phrases are revenue, sales, and demand

where these phrases are different but closely related. Sales and demand predictions can be close proxies for revenue. If the price of a good or service is not flexible, the revenue is always a fixed portion of total sales. The relationship between demand and revenue is more complicated than sales. In some cases, the product is highly differentiated in the market, and the firm produces and sells its products based on the forecasted demand. In this case, the revenue becomes a portion of the total sales and can be a proxy for revenue.

Articles predict revenue when there are many options to consider for pricing, demand, and factors affecting them. There are options to change the cost of the product or service such as adjusting the budget of advertisement, including or excluding some features in the products, for example, extra RAM or a better processor in a laptop. The standard approach to predict revenue was statistics, econometrics, and mathematical modeling for each case.

Among several approaches in the published articles. Data mining and machine learning techniques like associate rule mining, support vector machines, neural networks, and random forests used in more recent articles and when the data is significantly large, which usually comes from social media. In predicting revenue generated by selling goods or services, econometric models are more widespread in the earlier article, and there seems to be a shift toward using machine learning in recent ones.

Recently, with advances in machine learning techniques and accessibility of social media data, a new trend of revenue prediction articles has appeared that almost exclusively work on predicting the box office value of movies before their openings in the cinemas. This approach added to the classic methods of revenue prediction in two ways. First is the use of social media data and sentiment analysis, and the second is the use of machine learning instead of individual models or simple econometrics methods.

The use of machine learning created a new standard in this field. Researchers focus more on combining different methods to take advantage of the benefits of multiple models. It is now common to first classify the date and then run different models on each class. This method is by itself is a critical improvement over the old approach that limits the study to just a portion of data.

Most of the times, researchers use all the features they have in the data. Using all of the features in data is not always feasible. For example, in predicting revenue of movies, mainly the movie's box-office revenue, many researchers use social media contents and reviews written by users to predict the box-office revenue which has an enormous number of feature and using all of them is not feasible. Also, there is not enough correlation between some features and revenue to justify using many features. Feature selection is the approach to find which features are good predictors and limit the feature set to them. It is, however, possible to limit the feature set to a custom set which will help with the interpretability of the model but usually decreases accuracy.

The rest of this chapter systematically explore the literature to reveal insight on publication trends, area of application in revenue prediction, the technique of prediction, and the evaluation criteria.

2.1 Survey Methodology

This section examines the published papers by the year of publication and classifies them by year of publication. Another categorization is the area of application which clarifies the type of revenue in the literature, either being tax, goods, services or other applications. The prediction technique is classified separately to highlight popular methods. Similar to the techniques, the evaluation criteria are also categorized to find essential metrics for comparison between techniques.

The initial search started by collecting all the papers in major digital computer science databases. The databases that this survey has covered are:

- jstore
- Science Direct
- Springer
- IEEEXplore

Our focus is to find the articles that try to predict the generated revenue by the data collected from customers (like social media data and IP logs) therefore we eliminate the articles that work solely on demand forecasting or revenue management.

The initial search found 104 papers by searching for "Revenue prediction", "revenue forecasting", and "Revenue prediction" OR "revenue forecasting" in all fields. These papers carefully examined and only the papers that predict revenue for an organization by using data from customers were selected.

However, the initial search includes all the papers published in academic journals and conference proceedings that are subjected to peer review and have higher values.

After careful review of the selected papers, 59 papers met the criteria stated above. The next part of the survey categorizes these papers into the area of application, prediction technique, and evaluation criteria.

2.2 Classification Method

The selected papers categorized based on their year of publication to give us more insight into the publication trends. The classification framework consists of the area of application, prediction technique, and the evaluation criteria. Articles fall into six categories of Tax, goods, services, energy, movies, and others based on their area of application. The prediction techniques consist of eleven categories being regression

analysis, time-series analysis, mathematical modeling, associate rule mining, support vector regression, neural networks, random forests, factorization machines, mechanism design, and gene expression programming.

The evaluation criteria categorized into thirteen categories being precision, recall, F1 measure, mean classification rate, mean squared error, root mean square error, mean absolute error, relative absolute error, percentage error, R-squared, Pearson correlation coefficient, Spearman's rank correlation coefficient, and Akaike information criterion.

2.2.1 Area of Application

This section categorizes the articles based on their area of application. Movies category contains the papers that predict the revenue of a movie, whether Box-office, total sale, or the first week sales. Articles categorized under Energy that predict revenue generated by selling an energy source. Tax is the category that contains government revenue from all kinds of different taxes. Goods contains articles that predict the revenue from selling physical goods or virtual goods like software.

Any activity that does not transfer a good from seller to buyer considered as Service in our categorization system. The articles that do not fit in any of the groups mentioned above put into the Other category.

2.2.2 Prediction Technique

Prediction techniques in reviewed papers fall into different categories. Regression analysis is a well known statistical process for estimating the relationship between variables. There are many variations of this technique and all the papers using any of these methods put under "Regression analysis" group. Another well known statistical technique is the time series analysis that predicts the revenue using past data.

This method also has many variations and all the articles using one of these methods put under the "Time series analysis" category. In addition to these two categories, some mathematical models developed for a specific problem. These articles put under the category of "Mathematical modeling" to capture the effectiveness of these models. Because of the high customization of these models, they are easy to interpret and usually yield an accurate prediction, the downside of this approach is that mathematical modeling is sophisticated and need a good knowledge of the topic.

One of the standard machine learning techniques, especially in sentiment analysis from social media is Associate rule mining (ARM). ARM is a method for discovering relationships between variables in large data-sets. There are several extensions and algorithms to apply ARM to data, so all the papers using one of them put under the "ARM" category.

Support vector regressions machine (SVR) [7] and Least-squares support-vector machines (LS-SVM) [46] are machine learning techniques for data and regression analysis. They are both based on support vector machines and have many implementations to the point that it is hard to distinguish them from each other. This close similarity in concept and implementation causes all the papers using any of these models or their variation to fall in the "SVR" category. Factorization Machines (FM) are models combining support vector machines with factorization models. These models share the essential characteristics of both models [39] and papers using them fall into the "FM" category. The last Machine learning technique is the Random Forests (RF). RF is capable of classification and regression, and it is an enhancement of decision trees. RF also has many implementations and variations and the papers using any of these methods put in the "RF" category.

A common approach for a better prediction is to combine different methods. This method mostly used to improve prediction accuracy. Articles that use more than one

model in combination grouped under "Hybrid category. Two categories of mechanism design "MD" and gene expression programming "GEP" used less often than the other methods but not least important. Mechanism design purpose economic mechanisms or incentives toward desired objectives and GEP is an evolutionary algorithm for creating models. These models learn and adapt by mutating their sizes, shapes, and composition.

2.2.3 Evaluation Criteria

Different modeling techniques use different criteria to evaluate their performance. Reviewed articles use thirteen different evaluation criteria to illustrate the performance of their model. Precision, Recall, and F1-measure are metrics that commonly used together for evaluating the classification. "Precision" or positive predictive value is a fraction of relevant instances among the retrieved instances. The probability of detection, sensitivity, or "Recall" is the fraction of relevant retrieved instances over the total amount of relevant instances. "F1" measure is the harmonic mean of precision and recall. Another classification metric is the Mean Classification Rate (MCR), which is the ratio of the correctly classified observation.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{2.1}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{2.2}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{2.3}$$

$$MCR = \frac{TruePositive + TrueNegative}{TotalPopulation} \tag{2.4}$$

Other metrics for continues variables are Mean Squared Error (MSE) or mean squared deviation, t take the average of squared errors. Root Mean Square Error (RMSE) measures the standard deviation of residuals. Mean absolute error (MAE) measures the difference between predicted values and actual values. Relative Absolute Error (RAE) is the errors normalized by the total absolute errors and can capture the magnitude of the difference. Mean Absolute Percentage Error (MAPE) is the average percentage difference of the predicted values from the actual values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{2.5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{2.6}$$

$$RAE = |\frac{Y_i - \hat{Y}_i}{Y_i}| \tag{2.7}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i| \times 100\% \tag{2.8}$$

Where $\hat{Y}_i$ is predicted value for Y_i and Y_i is the actual value.

R-squared and Akaike information criterion (AIC) evaluate the performance of the model as a whole. R-squared is the coefficient of determination and measures what portion of the standard deviation in the dependent variable is determined by independent variables. AIC is a measure for model selection and estimates the relative

quality or information loss of a model.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2} \tag{2.9}$$

Where $\hat{Y}_i$ is predicted value for Y_i and $\bar{Y}$ is the mean of Y for i=1 to n.

$$AIC = 2k - 2Ln(\hat{L}) \tag{2.10}$$

Where k is the number of estimated parameters in the model and $\hat{L}$ is the maximum value of the likelihood function.

Correlation metrics usually used at the feature level for confirming the correlation of a feature with the prediction. Pearson correlation coefficient (PCC) is a measure of linear correlation between two variables and used to evaluate the correlation between predicted values and the actual values. Spearman's rank correlation coefficient (SRCC) is similar to PCC but assessed if the relationship in monotonic or not.

$$PCC_{xy} = \rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{2.11}$$

$$SRCC_s = \rho_{\mathrm{rg}_X,\mathrm{rg}_Y} = \frac{\mathrm{cov}(\mathrm{rg}_X,\mathrm{rg}_Y)}{\sigma_{\mathrm{rg}_X}\sigma_{\mathrm{rg}_Y}} \tag{2.12}$$

where $\mathrm{cov}(\mathrm{rg}_X,\mathrm{rg}_Y)$ is the covariance of the rank variable and $\sigma_{\mathrm{rg}_X}\sigma_{\mathrm{rg}_X}$ and $\sigma_{\mathrm{rg}_Y}\sigma_{\mathrm{rg}_Y}$ are the standard deviations of the rank variables.

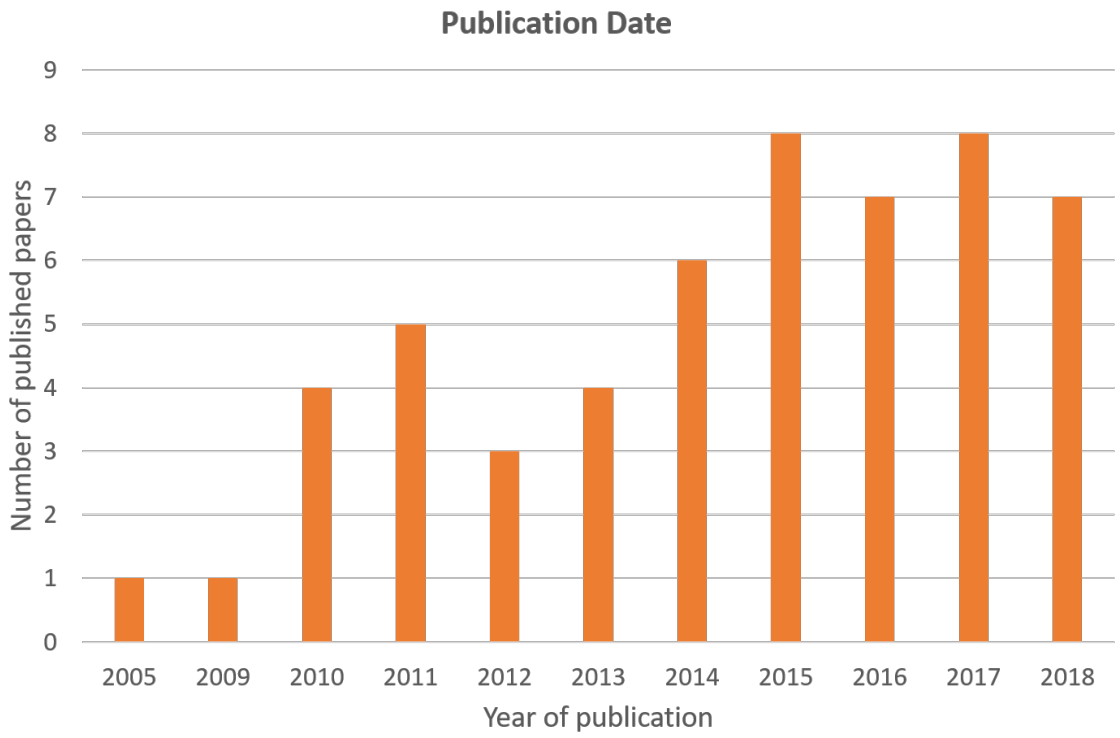

Figure 2.1: Distribution of research papers by year of publication

2.3 Classification Results

2.3.1 Publication Trend

The distribution of research papers by year of publication is presented in Figure 2.1. Research papers published between 2005 and 2018 and the number of published articles increased from 2009 to 2010 and remained the same until 2013. The revenue in the articles before 2013 is mainly in the area of goods, tax, and services, and there is a rise in the number of publications after 2013.

This increase corresponds mainly to the articles that predict the box-office revenue using various techniques statistical and machine learning techniques. With easy access to social media data and sentiment analysis for predicting the revenue of movies, the number of published articles increased and remained the same.

2.3.2 Area of Application

Movies Box-office revenues are predicted using social media contents in [29, 8, 18, 44, 48, 32]. However, data concerning movies are not limited to those in social media. There are other types of data, including ratings, and reviews and [53] used some of these online resources (search engines, video sites, and WeChat) for their prediction. Some of these important resources are IMDB , BoxOfficeMojo , Douban , Shanghai Media and Entertainment Group and FilmSource database.

Depending on the modeling approach, online resources used differently for movie revenue prediction either by itself or merged with other resources. [62, 38, 61, 54, 31, 30] used only one resource while [16, 40] combined two databases for their study. Data is also not limited only to online resources. For example, [6] gathered information from 167 participants of their experiments and [15] used the information provided by 68 Chinese theatres.

Goods divide into two categories: physical and virtual goods. Examples of physical goods are food, goods in supermarkets, furniture, and automotive that work of [50, 49, 45, 13, 9] studied. Virtual goods are the ones that exchanged and used virtually such as software. [1, 23] explored these kinds of goods.

Services are transactions that do not transfer any good (physical or virtual) from seller to buyers. Example of services are: online advertising [3, 11], tourism [26], music concerts [24], taxis [58], municipal services [17, 43] rentals [41], parking [14] and etc. [35, 52]

Government tax revenue literature explore tax revenue for different countries. For example, [51, 60, 57, 28] China, [5] Germany, [33] Indonesia and [22] OECD countries. Also, [12, 55] worked on government revenues in without any particular counrty.

Work of [25, 10, 27, 36] focus around the data on electricity and power plants so

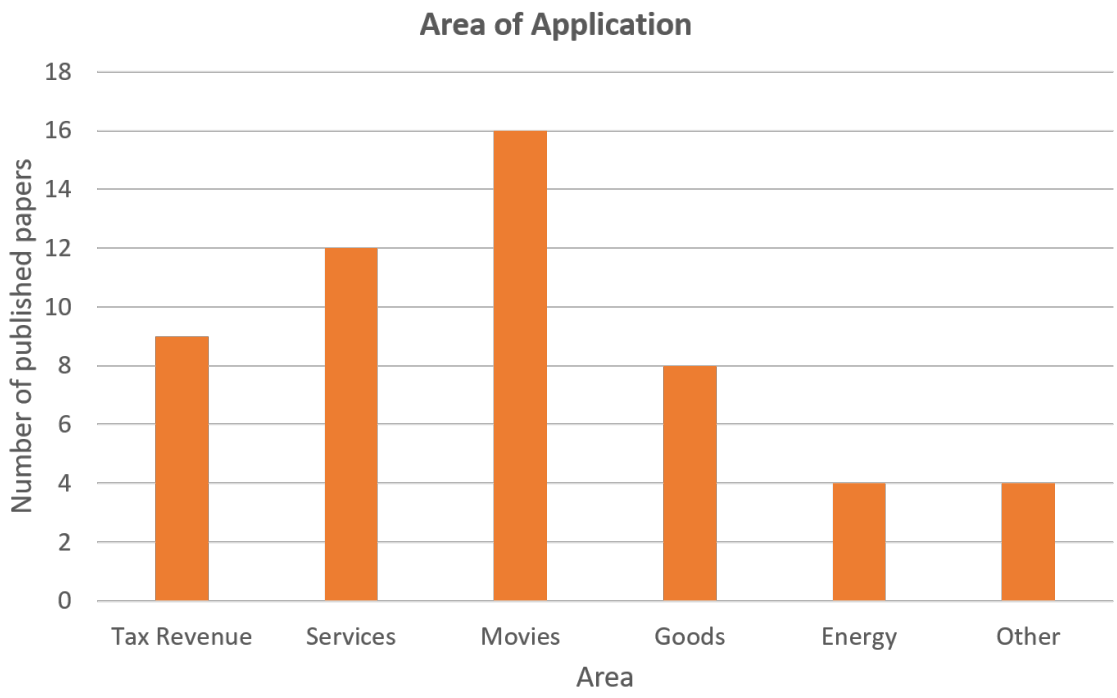

Figure 2.2: Distribution of research papers by their application area

fall under energy. Other articles that do not belong to any of the categories above are categorized under others ([21, 34, 20, 4]).

2.3.3 Prediction Technique

Figure 2.3 shows the distribution of techniques used for prediction. Regression analysis is the most practiced method for prediction in the reviewed papers. [29, 5, 32, 53, 52, 38, 22, 12, 16] used linear regression and [31] also considered the interactions between variables in their linear regression models moreover [35] included polynomial terms in their regression analysis. Other types of regression analysis utilized less than linear regression in the reviewed papers, including Multinomial Logit in [9] and Gaussian copula in [8].

The study of [29, 13, 21, 58] used ordinary SVR. The variation of SVR utilized in

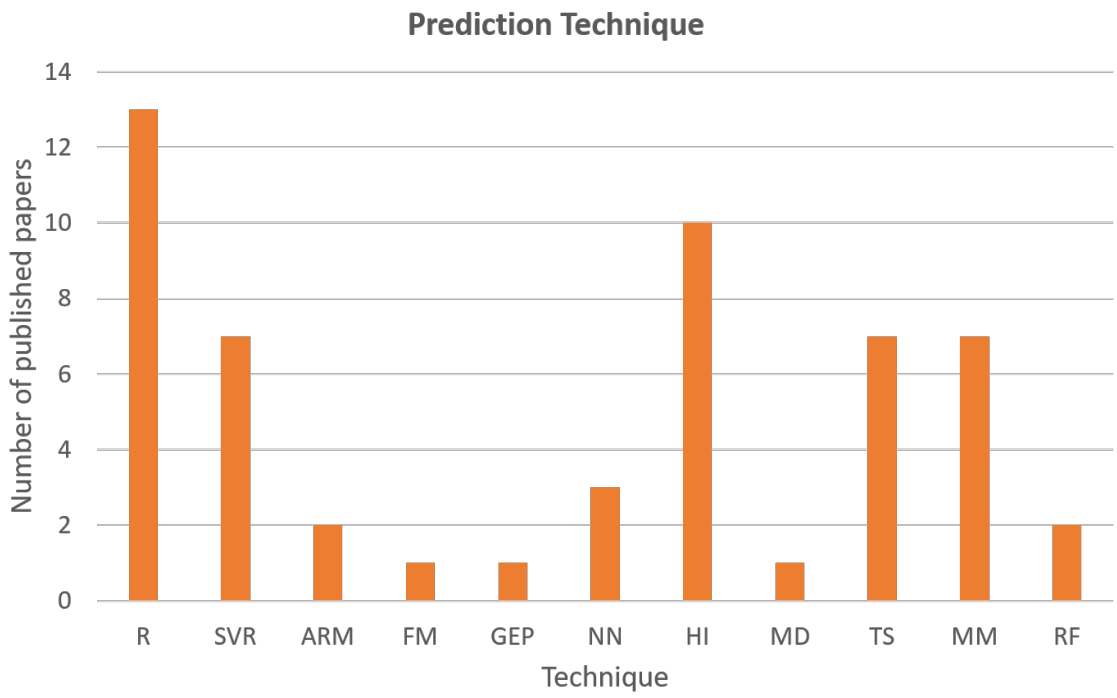

Figure 2.3: Distribution of research papers by prediction technique

some papers. For example, [51] used partial least square SVM, and [60] performed gray correlation analysis using SVR. Another variation of SVR is sequential mining optimization that [20] used in their study.

The techniques used in Time series analysis are ARIMAmodels in work of [43, 56, 1], Grey Forecasting in the study of [26, 28, 27], and Transfer Entropy in the article of [34].

Mathematical modeling used in the work of [24, 25, 3, 10, 14, 11, 23]. These articles formulated the prediction problem mathematically first and only used data to support their claims. There are situations that there are specific models for a problem that usually performs better or easily interpretable. Examples of these approaches are work of [54] with factorization machine, [36] with gene expression programming and [6] with mechanism design.

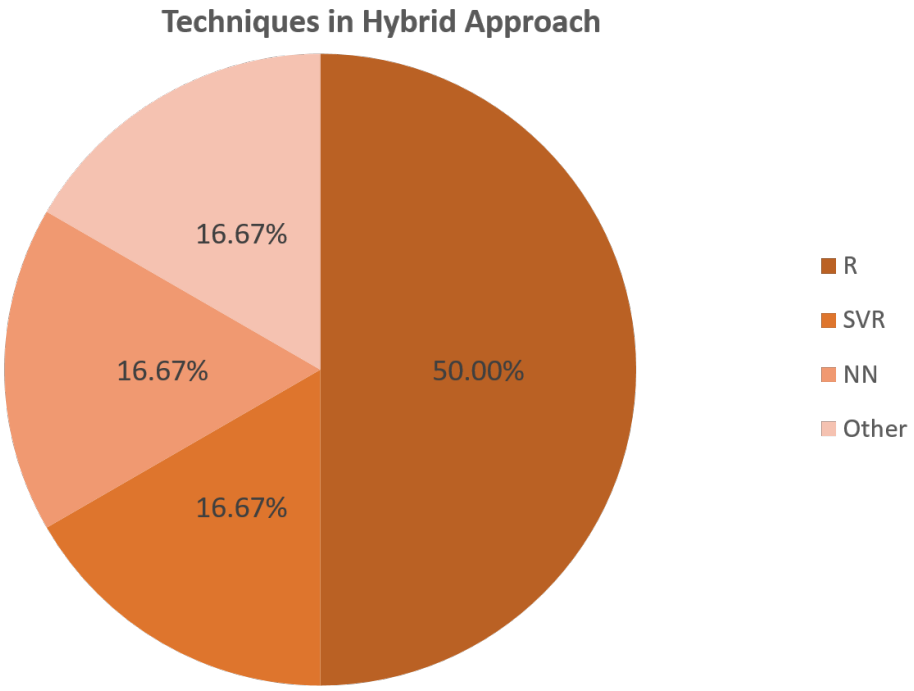

Figure 2.4: Percentage of models used in Hybrid approach for prediction

Associate Rule Mining is used to find the relationship between variables or categorization, and these relationships or categories are useful to predict the generated revenue. Work of[49, 50] demonstrates the use of ARM in prediction. Other methods of prediction are Neural Networks in [62, 41, 33] and Random Forests in [4, 15].

Another approach of prediction is combining two or more techniques to improve accuracy. The standard procedure is usually feeding the results of one model to another. Each of the models in the combination serves a different purpose. The accuracy in one of the models (usually the last one) is the target for the other therefore the first models generate results that work very good in the second model, the first model is responsible for classification or evolution. Work of [45] used Genetic Algorithm in addition to Neural Networks, and [55] used SVR plus Learning Vector Quantization to improve their results.

The other group that consists of the work of [17, 18, 44, 48, 40, 57, 61, 30], used a classification technique such as K-Means, SVM to first find possible categories and then predict each group separately by another method such as linear regression or SVR. Linear regression is the most common model followed by SVR and Neural Networks as the last model in a hybrid approach. The distribution of the models used in the hybrid approach is presented in Figure 2.4.

2.3.4 Evaluation Criteria

Figure 2.5 shows the distribution of research papers by their evaluation criteria. The relative absolute error had the most application in the reviewed articles and used in four different ways. [27, 26, 55, 35, 60] used RAE to compare the result of their final model with the baseline approach, and [58, 54, 20, 16] used RAE to evaluate the model generated results with actual values. [9, 31] used RAE as a model selection criteria to evaluate different methods, [5] used RAE to find the difference between monthly forecast results, and [32] used it to evaluate the predicted values for different teams.

The mean absolute percentage error is the second commonly used criteria. [13, 48, 30] used MAPE to compare different scenarios, and [43] used it to select the appropriate ARIMA model, also [55] used MAPE to compare their model with a baseline; moreover, [1, 51, 36, 28] used MAPE to evaluate the performances of different models.

R-squared had various usage in reviewed articles. [29] made a comparison between the proposed model with a baseline model using R-squared. [4] evaluated the performance of their model using real data, and [21, 40, 53, 36, 52] used R-squared to compare different model preferences or different data. AIC did not have diverse usages, [18, 12] used AIC to rank different models.

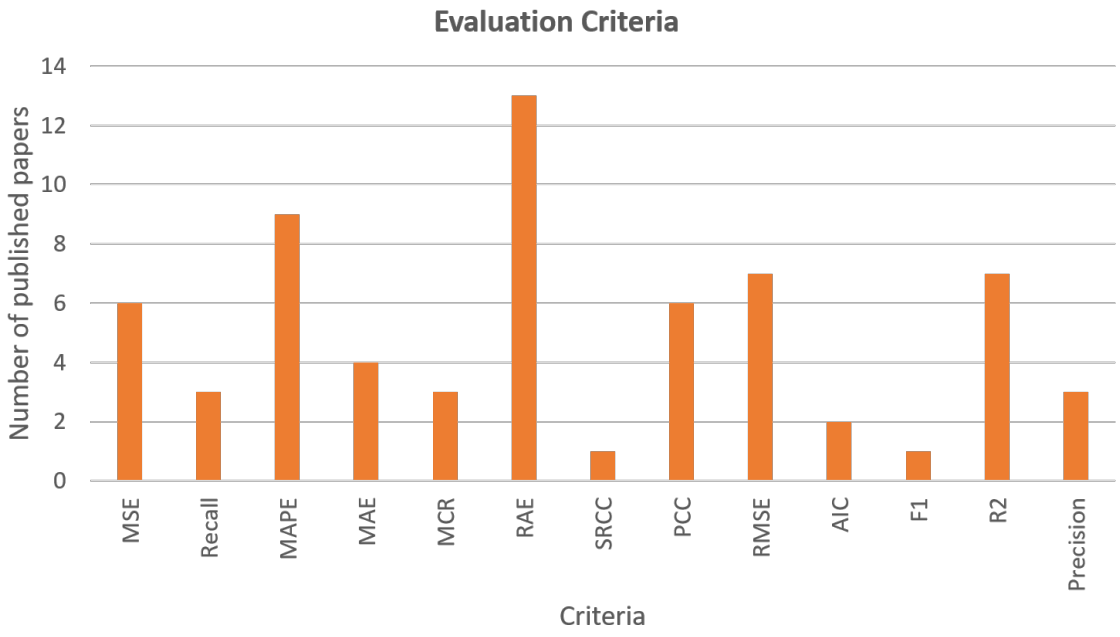

Figure 2.5: Distribution of research papers by the evaluation criteria

The squared errors also had multiple usages. [8, 1, 18, 35] used mean absolute error to evaluate the results of different models. [44, 60, 11, 36, 22, 6, 12] used the root mean squared error to evaluate their results with real data. [21, 17, 33] used MSE to evaluate true values against predicted ones and [1, 51, 41] used MSE to evaluate the performance of different models.

The primary use of Correlation coefficients was model ranking. [8, 18, 38] used Pearson's correlation coefficient and [56] used Spearman ranked correlation coefficient to evaluate the difference of the performance between various models. also, [53, 23] used PCC to show the relationship between predicted values and actual values.

The classification measures primary usage was comparing different models. [49] used precision to evaluate the accuracy of different models. [50, 34] used precision and recall for model selection, and [29] used F1 measure in addition to precision and recall for feature selection.

Similarly, [45] used mean classification rate to evaluate different models and [62, 61, 15] used a similar measure (average present hit rate) to evaluate different models.

2.4 Chapter Conclusion

The results of the literature clearly show that revenue prediction is a fertile possible future research. The increasing number of publication in recent years and availability of data, especially from social media and logs, along with new prediction methods shows the promising future of this field.

This survey clarifies the application areas of revenue prediction. Movies and services are the areas with the most publications. This result is reasonable because movies and services had easily accessible data for a longer time. The availability of data from online stores that sell goods opens a possible future direction. However, data from tax revenue and energy do not seem to be publicly available.

The most used prediction techniques are regression analysis, hybrid methods, and SVR. Regression analysis is straightforward and usually generates acceptable results, and it is considered a baseline for many studies. Combining different methods to create more accurate predictions is always a possible future research opportunity. Some Machine learning techniques only used in Movies and service revenue. Applying these techniques to different areas is another potential future work.

None of the reviewed articles using machine learning techniques limited the features or tried to interpret their predictions. The algorithm selects the features automatically based on their correlation with the target variable. Future research can explore limiting the features to a more meaningful set in order to make them interpretable. Additionally, new data always create an opportunity for research.

This section clearly shows the limitations, challenges, and opportunities in the

revenue prediction that was not available before. It presents the trends, application, and techniques of available literature, and It serves as a starting point for researchers attracted to revenue prediction.

Chapter 3

Methodology

This section explains the techniques and evaluation criteria used for the revenue prediction. Also, it explains the implementation and provide some descriptive analytics about the dataset.

3.1 Regression Analysis

Regression analysis regularly used for prediction and forecasting. This thesis uses linear regression, which treats the Revenue from each customer as a dependent variable, and the features are independent. This model does not consider interactions between the independent variable, and it is used here as a baseline for comparison. The predicted model is as shown in Equation 3.1.

$$Log(Revenue_{fullVisitorId}) = \alpha + \sum_{i \in \Omega} \beta_i x_i + \epsilon_i \tag{3.1}$$

Where Ω is the set of selected features and x_i is i^{th} feature.

3.2 Neural Networks

Artificial Neural Networks (NN) is a framework to recognize the pattern in data. There are many methods to train the neural network. This thesis uses back-propagation

that is one of the most established methods to calculate the gradient of the loss function concerning the weights in an NN. The revenue function then can be calculated as in Equation 3.2.

$$Log(Revenue_{fullVisitorId})_{n+1} = Log(Revenue_{fullVisitorId})_n - \gamma \nabla F(x_i),\ i \in \Omega \quad (3.2)$$

Where n is the number of iterations, γ is the step size and F is a convex function of features $x_i,\ i \in \Omega$.

3.3 Convolutional Neural Network

Convolutional Neural Network (CNN) is a class of deep neural networks. In this class, each neuron computes an output value by applying some function to the input values coming from the receptive field in the previous layer. This function is specified by a vector of weights (plus bias). Learning continues by making incremental adjustments to the biases and weights. The vector containing weights and bias is called a filter and represents some feature of the input. A distinguishing feature of CNNs is that many neurons share the same filter. The convolutional layer consists of a set of kernels, which have a small receptive field but extend through the full depth of the input volume. Figure 3.1 shows a high-level description of CNN.

3.4 Least-squares support-vector machine

Least-squares support-vector machines (LS-SVM similar to SVR) are a version of support vector machines and used for data analysis and pattern recognition. This prediction used a Gaussian kernel for training 80% of the data and used the rest for testing. For details of computation and more explanation of the model, see [46]. The

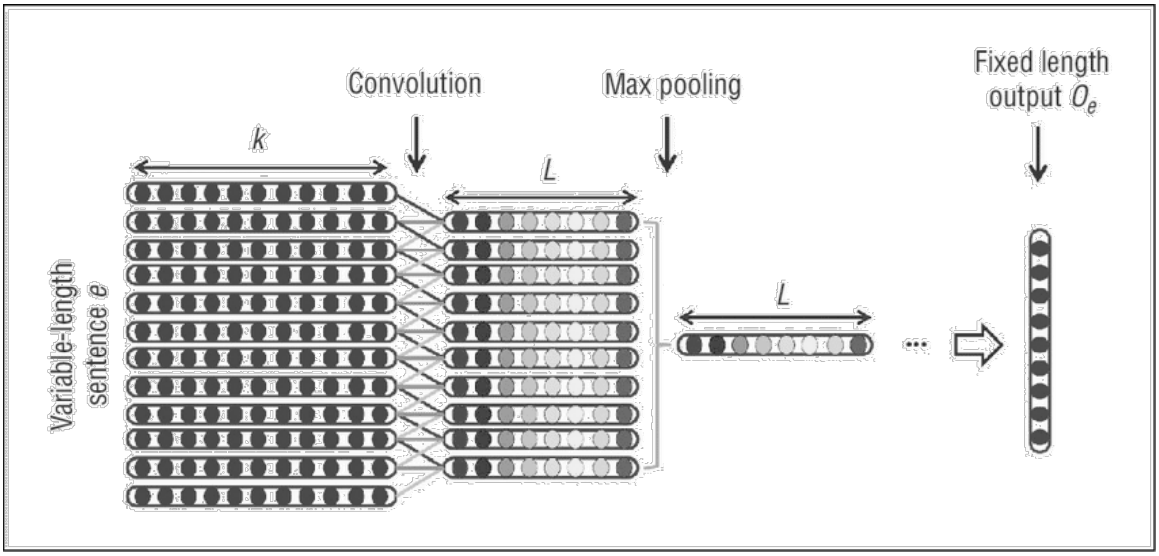

Figure 3.1: High-level architecture of a CNN [59]

method minimized the error as in Equation 3.3.

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i e_i)^2 = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left(y_i - (w^T \phi(x_i) + b)\right)^2 \tag{3.3}$$

where $e_i = y_i - (w^T \phi(x_i) + b)$, $k(\vec{x}_i, \vec{x}_j) = \phi(\vec{x}_i) \cdot \phi(\vec{x}_j)$ is the kernel function, and $\vec{w} = \sum_{i=1}^{n} c_i y_i \phi(\vec{x}_i)$.

3.5 LS-SVM and Boosting

The primary purpose of boosting is reducing variance and bias. Boosting can turn weak learners to strong learners by repetitively calculating new weights and adjusting the feature's weights. This part used the LS-SVM model in section 3.4 for training. The mean loss function, coming from the probability of a sample drawing with LS-SMV, can update the weights in the original LS-SVM model. For details of this method see [42].

3.6 Experiment

3.6.1 Dataset

The data originates from the Google Analytic demo account, and it contains the Google Merchandise Store customer database. It contains 903,653 observations and 55 features in JSON format. These features arranged into 12 groups and the information about the geography stored under geoNetwork. For details of the headers and groups see Table 3.1.

Most of the customers of the store do not make a purchase, and customers who make a purchase will not purchase during all their visits. This pattern is visible in the data, and a large number of visits have zero revenue (77.8%). Most of the visitors are from the Americas, Asia, and Europe furthermore most of the non zero revenues are from the Americas. Africa has the most value of mean revenue, meaning that its low number of visits are most likely to end up in a high revenue from the store. Figure 3.2 shows the distribution of user based on their continents.

The data is used as-is, and due to its nature, there is no need for cleaning or making any adjustments. However, to run the regression models, the JSON file needs to be flattened.

Processing this file creates its challenges. It does not fit into memory and should be split into separate parts. The processing time is heavily related to the number of observations in each part that goes through the model. The next section will explain the solution to these problems.

3.6.2 Coding and Implementation

The models in the previous sections implemented using Python programing language. All Models implementations utilize SKlearn libraryin python along with standard

Table 3.1: List of the headers in the Data. Note that because the data is in JSON format, Each row might present several features. There are 55 features in total.

DataField	**Explanation**
fullVisitorId	A unique identifier for each user of the Google Merchandise Store.
channelGrouping	The channel via which the user came to the Store.
date	The date on which the user visited the Store.
device	The specifications for the device used to access the Store.
geoNetwork	This section contains information about the geography of the user.
sessionId	A unique identifier for this visit to the store.
socialEngagementType	Engagement type, either "Socially Engaged" or "Not Socially Engaged".
totals	This section contains aggregate values across the session.
trafficSource	This section contains information about the Traffic Source from which the session originated.
visitId	An identifier for this session. This is only unique to the user. For a completely unique ID, you should use a combination of fullVisitorId and visitId.
visitNumber	The session number for this user. If this is the first session, then this is set to 1.
visitStartTime	The timestamp (expressed as POSIX time).

packages like pandas and numpy. The implementations in SKlearn are highly efficient both in terms of memory and processor; therefore, eliminate the need to implement any of the models, especially when models are highly recognized and commonly used. Coding these models are highly depreciated.

One of the difficulties of using big data is processing power and memory, which

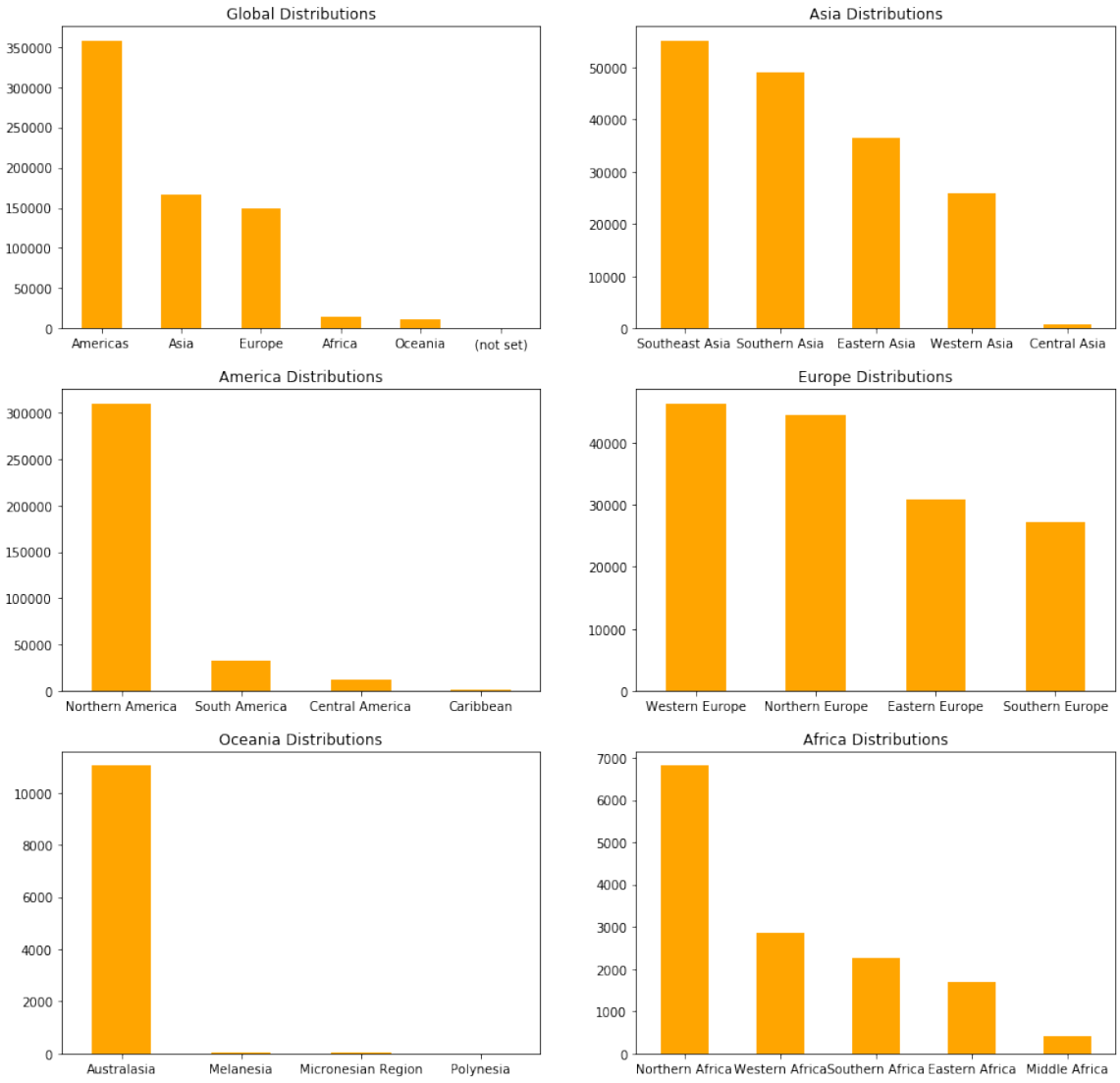

Figure 3.2: Distribution of users in 5 continents.

were limited in this case. Due to the size of this dataset, it is not possible to directly store it in the memory to run models. Instead, another approach is selected, which breaks the file into chunks. The data is divided into separate batches randomly, and each batch fitted to the model. Figure 3.3 shows the flow of this process.

The number of batches is directly related to the amount of available memory, and in the case of this thesis, two chunks can solve the problem of the memory but not

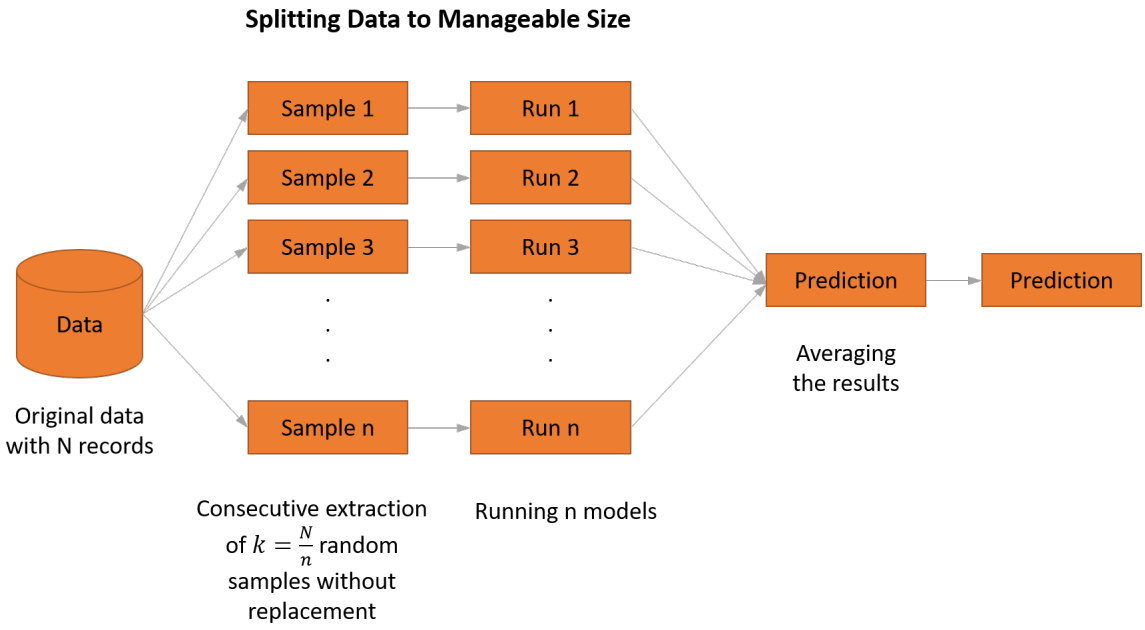

Figure 3.3: Batch process for limited memory

the computation time. Splitting the model into ten parts seem to capture the best for memory and the processing time.

3.6.3 Evaluation Metrics

Choosing the evaluation metric is difficult because there is no single metric that can capture all aspects of a model. Two possible candidates are Relative Absolute Error (RAE), and Root Mean Squared Error (RMSE). Both of these metrics determine the prediction error of the model, hence are measures of accuracy. RMSE is more sensitive to the difference between actual values and predicted value (sensitive to outliers). On the other hand, RAE differentiates between the prediction of low or high values, and it is useful when the distribution of the actual values is broad.

Relative Absolute Error (RAE) is beneficial in four different situations. As it is evident from the literature, it is a good tool for evaluating the prediction results with actual values.[27, 26, 55, 35, 60] used RAE to compare the result of their final model

with the baseline approach. [58, 54, 20, 16, 5, 32] used RAE to evaluate the model generated results with actual values. [9, 31] used it for model selection criteria to evaluate different methods.

$$RAE = |\frac{Y_i - \hat{Y}_i}{Y_i}| \tag{3.4}$$

RAE is a helpful metric since it can create a common ground for comparing different models, and it is especially helpful since the range of the firm's revenue from different customers is wide. RAE makes it possible to account for this range and does not penalize models only based on their prediction for extremely low or high values.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{3.5}$$

The RMSE combines the magnitudes of the errors in predictions of each observation into a single measure that implies the prediction power. RMSE is non-negative, and a value of zero shows a perfect fit to the data. Low values of RMSE show better prediction power, but all models in comparison should use the same data because RMSE is dependent on the scale of the values in data.

Chapter 4

Results and Discussion

Figure 4.1 and 4.2 show the comparison of 5 different modelings techniques based on the calculated RAE and RMSE. The results from the regression analysis generate the highest errors both in terms of RAE and RMSE. Although LS-SVM should have less error than the Regression, LS-SVM results are not significantly better. RAE and RMSE of Regression is 31.093% and 25.081% higher than LS-SVM. Both of these models use the least squared method for minimizing the prediction error.

Results of neural networks with back-propagation is a significant improvement to both LS-SVM and Regression. RAE and RMSE are 49.573% and 38.238% lower than the Regression and 33.894% 22.747% lower than LS-SVM. This improvement is mainly the result of the optimization method.

The results of convolutional neural networks are 33.051% lower RAE and 23.049% reduction in RMSE respect to neural networks. The LS-SVM and Boosting had the least error in terms of RAE and RMSE. These values are lower by 75.000% for RAE and 73.841% for RMSE respect to the Regression. These values are 25.949% lower in RAE and 44.959% lower in RMSE respect to CNN.

Between the compared approaches, LS-SVM with Boosting achieved the least error (0.117 for RAE and 0.482 RMSE) and considered the most accurate model. CNN achieved a RAE of 0.158 and RMSE of 0.877 and ranked second.

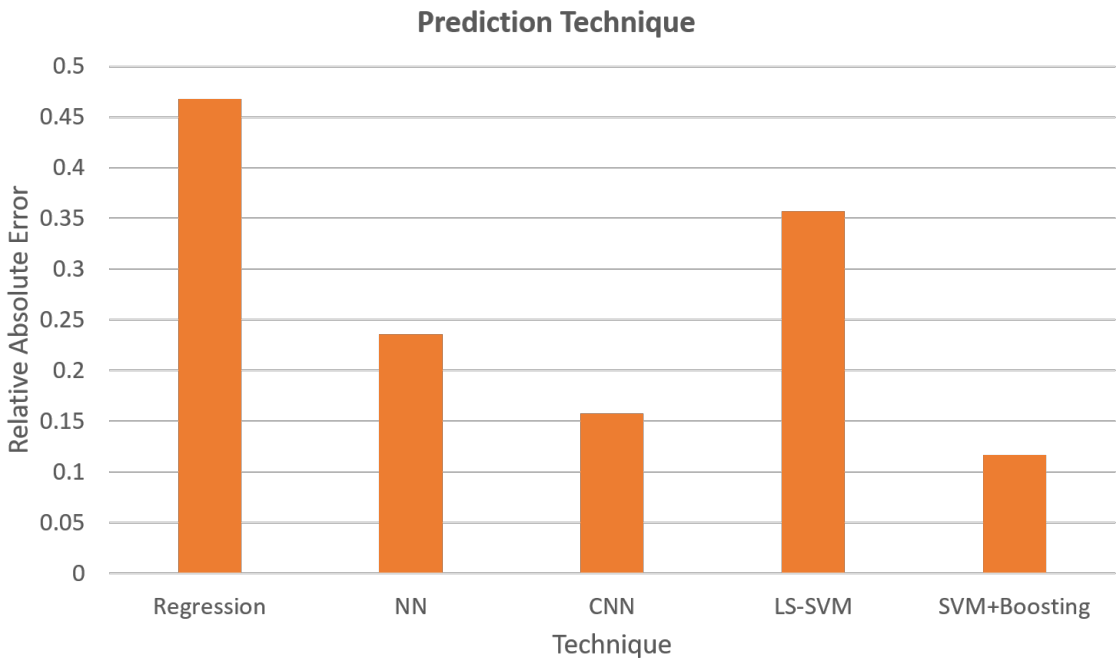

Figure 4.1: Comparison of the models base on the RAE on the test data

There are some benefits to both models. In terms of accuracy, LS-SVM produces a more accurate prediction. On the other hand, the implementation of CNN is much simpler. Figures 4.1 and 4.2 show the ranking and percentage of improvement to the results of Regression.

Table 4.1: Comparison of models and their rankings based on the percentage improvement of Regression RAE

Model	**Rank**	**RAE**	**Improvement**
Regression	0	0.468	0.00%
LS-SVM+Boosting	1	0.117	75.00%
CNN	2	0.158	66.24%
NN	3	0.236	49.57%
LS-SVM	4	0.357	23.72%

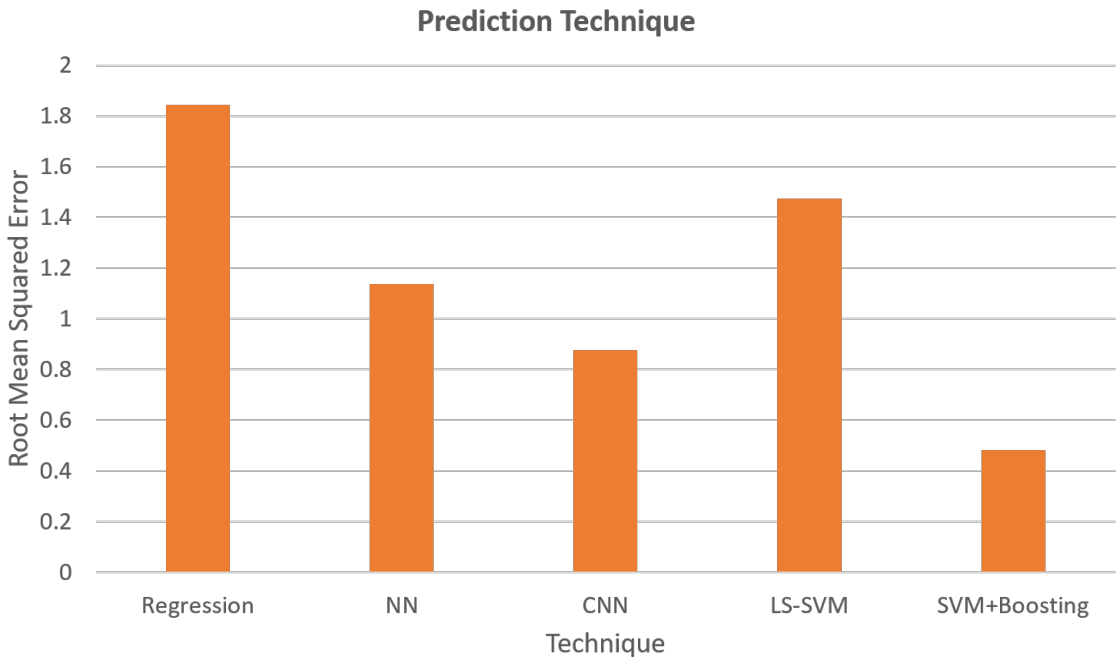

Figure 4.2: Comparison of the models base on the RMSE on the test data

Table 4.2: Comparison of models and their rankings based on the percentage improvement of Regression RMSE

Model	Rank	RMSE	Improvement
Regression	0	1.844	0.00%
LS-SVM+Boosting	1	0.482	73.84%
CNN	2	0.877	52.47%
NN	3	1.139	38.24%
LS-SVM	4	1.475	20.05%

Chapter 5

Conclusion and Future Research

The results of this study show that geographical information can be a good predictor for firms revenue. The comparison between Regression, convolutional neural networks, LS-SVM, Neural Networks with back propagation, and LS-SVM with Boosting show that the best model in terms of accuracy is the ensemble of LS-SVM and boosting with RAE of 0.117 or 88.3% and RMSE of 0.4824628.

One of the critical findings in this work is the effective modeling of features with a strong to weak correlation with the target value. Some models like CNN have a built-in mechanism to deal with this issue, and some models like LS-SVR need the help of other boosting mechanisms. It is possible to achieve an accurate prediction by mixing strong and weak learners.

Some straight forward future research opportunities are exploring different modeling approaches to obtain a better accuracy for prediction, working on city-level data to eliminate the differences between countries and exploring different application areas such as movies.

This work shows the relationship between the location of customers and the generated revenue for the firm. The cause of this relationship is complicated and cannot be linked directly to the location. Customers in an area might receive regular advertisements, or there is a local burst in positive reviews by local users or a new trend

in the area. Another possible explanation can be that the location of customers is a proxy of their income. Exploring the causality is a future line of research.

BIBLIOGRAPHY

[1] Al-Ghamdi, M., Chester, A. P., and Jarvis, S. A. Predictive and dynamic resource allocation for enterprise applications. In *Proceedings - 10th IEEE International Conference on Computer and Information Technology, CIT-2010, 7th IEEE International Conference on Embedded Software and Systems, ICESS-2010, ScalCom-2010* (2010), pp. 2776–2783.

[2] Arnott, R. J., and Stiglitz, J. E. Aggregate land rents, expenditure on public goods, and optimal city size. *The Quarterly Journal of Economics 93*, 4 (1979), 471–500.

[3] Bax, E., Kuratti, A., McAfee, P., and Romero, J. Comparing predicted prices in auctions for online advertising. *International Journal of Industrial Organization 30*, 1 (2012), 80–88.

[4] Benedict, S. Revenue oriented air quality prediction microservices for smart cities. In *2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2017* (2017), vol. 2017-Janua, pp. 437–442.

[5] Buettner, T., and Kauder, B. Political biases despite external expert participation? An empirical analysis of tax revenue forecasts in Germany. *Public Choice 164*, 3-4 (sep 2015), 287–307.

[6] Court, D., Gillen, B., McKenzie, J., and Plott, C. R. Two information aggregation mechanisms for predicting the opening weekend box office revenues of films: Boxoffice Prophecy and Guess of Guesses. *Economic Theory 65*, 1 (jan 2018), 25–54.

[7] Drucker, H., Burges, C. J., Kaufman, L., Smola, A. J., and Vapnik, V. Support vector regression machines. In *Advances in neural information processing systems* (1997), pp. 155–161.

[8] Duan, J., Ding, X., and Liu, T. A Gaussian copula regression model for movie box-office revenues prediction. *Science China Information Sciences 60*, 9 (sep 2017), 092103.

[9] Farias, V. F., Jagabathula, S., and Shah, D. A Nonparametric Approach to Modeling Choice with Limited Data. *Management Science 59*, 2 (2009), 305–322.

[10] Fisher, M., Whitacre, J., and Apt, J. A Simple Metric for Predicting Revenue from Electric Peak-Shaving and Optimal Battery Sizing. *Energy Technology 6*, 4 (2018), 649–657.

[11] France, S. L., Vaghefi, M. S., and Zhao, H. Characterizing viral videos: Methodology and applications. *Electronic Commerce Research and Applications 19* (sep 2016), 19–32.

[12] Freeman, J., Keith, J., Roberts, M., and Owens, A. The Effects of Revenue and Social Capital on Collective Governance: Implications for Political Complexity. *Cross-Cultural Research 52*, 4 (2018), 351–380.

[13] Golderzahi, V., and Pao, H. K. Understanding customers and their grouping via wifi sensing for business revenue forecasting. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10935 LNAI. Springer, Cham, jul 2018, pp. 56–71.

[14] Gongjun Yan, Weiming Yang, Rawat, D. B., and Olariu, S. SmartParking: A Secure and Intelligent Parking System. *IEEE Intelligent Transportation Systems Magazine 3*, 1 (2011), 18–30.

[15] Guo, Z., Zhang, X., and Hou, Y. Predicting box office receipts of movies with pruned random forest. In *International Conference on Neural Information Processing* (2015), Springer, pp. 55–62.

[16] Gupta, N., Abhinav, K. R., and Annappa, A. Fuzzy sentiment analysis on microblogs for movie revenue prediction. In *Proceedings - 2013 International Conference on Emerging Trends in Communication, Control, Signal Processing and Computing Applications, IEEE-C2SPCA 2013* (2013), pp. 1–4.

[17] Hájek, P., and Olej, V. Municipal Revenue Prediction by Support Vector Machine Ensembles 2 Parameters Design for Municipal Revenue Prediction. In *LATEST TRENDS on COMPUTERS* (2010), vol. 1, pp. 325–330.

[18] He, G., and Lee, S. Multi-model or single-model?: A study of movie box-office revenue prediction. In *Proceedings - 15th IEEE International Conference on Computer and Information Technology, CIT 2015, 14th IEEE International Conference on Ubiquitous Computing and Communications, IUCC 2015, 13th IEEE International Conference on Dependable, Autonomic and Se* (2015), pp. 321–325.

[19] Hoch, I. Income and city size. *Urban Studies 9*, 3 (1972), 299–328.

[20] Hsieh, W.-L., Hwang, S.-Y., Huang, H.-C., and Chang, S. Predicting company revenue trend using financial news. In *Pacific Asia Conference on Information Systems, PACIS 2016 - Proceedings* (2016).

[21] Jin, Y., Lin, C. Y., Matsuo, Y., and Ishizuka, M. Mining dynamic social networks from public news articles for company value prediction. *Social Network Analysis and Mining 2*, 3 (sep 2012), 217–228.

[22] Jochimsen, B., and Lehmann, R. On the political economy of national tax revenue forecasts: evidence from OECD countries. *Public Choice 170*, 3-4 (mar 2017), 211–230.

[23] Kannan, K., Goyal, M., and Jacob, G. T. Modeling the impact of review dynamics on utility value of a product. *Social Network Analysis and Mining 3*, 3 (sep 2013), 401–418.

[24] Kawahata, Y., Genda, E., and Ishii, A. Revenue prediction of music concerts using the mathematical model of hit phenomena. In *Proceedings - 2013 International Conference on Biometrics and Kansei Engineering, ICBAKE 2013* (2013), pp. 208–213.

[25] Kinsey, G. S. Spectrum sensitivity, energy yield, and revenue prediction of PV modules. In *IEEE Journal of Photovoltaics* (2015), vol. 5, pp. 258–262.

[26] Li, J. Application of improved grey GM(1,1) model in tourism revenues prediction. In *Proceedings of 2011 International Conference on Computer Science and Network Technology, ICCSNT 2011* (2011), vol. 2, pp. 800–802.

[27] Li, S., Ma, X., and Yang, C. A combined thermal power plant investment decision-making model based on intelligent fuzzy grey model and ito stochastic process and its application. *Energy 159* (sep 2018), 1102–1117.

[28] Liu, D., Zhang, R., and Li, J. Tax revenue combination forecast of Hebei province based on the IOWA operator. In *Proceedings - 4th International Joint Conference on Computational Sciences and Optimization, CSO 2011* (2011), pp. 516–519.

[29] Liu, T., Ding, X., Chen, Y., Chen, H., Guo, M., Liu, T., Ding, . X., Chen, Y., Chen, H., and Guo, M. Predicting movie Box-office revenues by exploiting large-scale social media content. *Multimedia Tools and Applications 75*, 3 (feb 2016), 1509–1528.

[30] Liu, Y., Yu, X., An, A., and Huang, X. Riding the tide of sentiment change: Sentiment analysis with evolving online reviews. *World Wide Web 16*, 4 (jul 2013), 477–496.

[31] Lovallo, D., Clarke, C., and Camerer, C. Robust analogizing and the outside view: Two empirical tests of case-based decision making. *Strategic Management Journal 33*, 5 (2012), 496–512.

[32] Lu, Y., Wang, F., and Maciejewski, R. Business intelligence from social media: A study from the VAST box office challenge. *IEEE Computer Graphics and Applications 34*, 5 (2014), 58–69.

[33] Lubis, F. A., and Albarda. Data partition and hidden neuron value formulation combination in neural network prediction model: Case study: Non-tax revenue prediction for Indonesian government unit. In *2018 International Conference on Information and Communications Technology, ICOIACT 2018* (2018), vol. 2018-Janua, pp. 879–884.

[34] McKenney, D., and White, T. Selecting transfer entropy thresholds for influence network prediction. *Social Network Analysis and Mining 7*, 1 (2017).

[35] Megahed, A., Yin, P., and Nezhad, H. R. M. An optimization approach to services sales forecasting in a multi-staged sales pipeline. In *Proceedings - 2016 IEEE International Conference on Services Computing, SCC 2016* (2016), pp. 713–719.

[36] Mousavi, S. M., Mostafavi, E. S., and Hosseinpour, F. Gene expression programming as a basis for new generation of electricity demand prediction models. *Computers and Industrial Engineering 74*, 1 (2014), 120–128.

[37] Mullainathan, S., and Spiess, J. Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives 31*, 2 (2017), 87–106.

[38] Okazaki, M., Nagasaka, R., and Miyata, R. Prediction of the box-office with publicly available information and web search volume. In *ICIIBMS 2015 - International Conference on Intelligent Informatics and Biomedical Sciences* (2016), pp. 418–419.

[39] Rendle, S. Factorization machines. In *2010 IEEE International Conference on Data Mining* (2010), IEEE, pp. 995–1000.

[40] Ruan, D. R., Liu, T., and Gao, K. Modelling on movie box-office prediction based on LFM algorithm. In *Proceedings of 2015 7th International Conference on Modelling, Identification and Control, ICMIC 2015* (2016), pp. 1–5.

[41] Sanjaya, C., Liana, M., and Widodo, A. Revenue prediction using artificial neural network. In *Proceedings - 2010 2nd International Conference on Advances in Computing, Control and Telecommunication Technologies, ACT 2010* (2010), pp. 97–99.

[42] Schapire, R. E. The strength of weak learnability. *Machine learning 5*, 2 (1990), 197–227.

[43] Shi, H. Y., Tsai, J. T., Ho, W. H., and Lee, K. T. Autoregressive integrated moving average model for long-term prediction of emergency department revenue and visitor volume. In *Proceedings - International Conference on Machine Learning and Cybernetics* (2011), vol. 3, pp. 979–982.

[44] Shim, S., and Pourhomayoun, M. Predicting movie market revenue using social media data. In *Proceedings - 2017 IEEE International Conference on Information Reuse and Integration, IRI 2017* (2017), vol. 2017-Janua, pp. 478–484.

[45] Stahlbock, R., and Crone, S. F. Evolutionary neural classification for evaluation of retail stores and decision support. In *Proceedings of the International Joint Conference on Neural Networks* (2005), vol. 3, pp. 1499–1504.

[46] Suykens, J. A., and Vandewalle, J. Least squares support vector machine classifiers. *Neural processing letters 9*, 3 (1999), 293–300.

[47] Sveikauskas, L. The productivity of cities. *The Quarterly Journal of Economics 89*, 3 (1975), 393–413.

[48] Tang, W. H., Yeh, M. Y., and Lee, A. J. Information diffusion among users on Facebook fan pages over time: Its impact on movie box office. In *DSAA 2014 - Proceedings of the 2014 IEEE International Conference on Data Science and Advanced Analytics* (2014), pp. 340–346.

[49] Weng, C. H. Discovering highly expected utility itemsets for revenue prediction. *Knowledge-Based Systems 104* (jul 2016), 39–51.

[50] Weng, C. H. Revenue prediction by mining frequent itemsets with customer analysis. *Engineering Applications of Artificial Intelligence 63* (aug 2017), 85–97.

[51] Wu, P. Particle swarm optimized partial least square support vector regression model for tax revenue prediction. *Journal of Chemical and Pharmaceutical Research 6*, 2 (2014), 38–46.

[52] Wu, Z., Chen, S., Wu, H., Cheng, Y., Dong, L., and Li, C. Measuring economic activity in China with mobile big data. *EPJ Data Science 6*, 1 (dec 2017), 29.

[53] Xiao, J., Yang, S. N., Li, X., Jin, H., and Chen, S. Quantifying domestic movie revenues using online resources in China. *International Journal of Simulation: Systems, Science and Technology 17*, 2 (2016), 4.1–4.9.

[54] Xu, Y., Tang, Q., Hou, L., and Li, M. Decision Model for Market of Performing Arts with Factorization Machine. *Journal of Shanghai Jiaotong University (Science) 23*, 1 (feb 2018), 74–84.

[55] Yang, H., He, J., and Jiang, F. Hybrid Genetic Algorithm and Support Vector Machine Performance in Public Fiscal Revenue Prediction. In *Chinese Control Conference, CCC* (2018), vol. 2018-July, pp. 4495–4499.

[56] Yang, L., Dimitrov, S., and Mantin, B. Forecasting sales of new virtual goods with the elo rating system. *Journal of Revenue and Pricing Management 13*, 6 (dec 2014), 457–469.

[57] Yu, Z., and Ji, H. Research of tax revenue intelligent forecast system. In *Proceedings - 2010 International Forum on Information Technology and Applications, IFITA 2010* (2010), vol. 3, pp. 112–114.

[58] Zhang, D., Sun, L., Li, B., Chen, C., Pan, G., Li, S., and Wu, Z. Understanding taxi service strategies from taxi GPS traces. *IEEE Transactions on Intelligent Transportation Systems 16*, 1 (2015), 123–135.

[59] Zhang, J., and Zong, C. Deep neural networks in machine translation: An overview. *IEEE Intelligent Systems 30* (09 2015), 16–25.

[60] Zhang, Y. Research on the model of tax revenue forecast of Jilin province based on gray correlation analysis. In *Proceedings - 2014 6th International Conference on Intelligent Human-Machine Systems and Cybernetics, IHMSC 2014* (2014), vol. 2, pp. 30–33.

[61] Zhang, Z., Chai, J., Li, B., Wang, Y., An, M., and Deng, Z. Movie Box Office Inteval Forecasting Based on CART. In *Proceedings - 2015 8th International Symposium on Computational Intelligence and Design, ISCID 2015* (2016), vol. 2, pp. 87–90.

[62] Zhou, Y., Zhang, L., and Yi, Z. Predicting movie box-office revenues using deep neural networks. *Neural Computing and Applications* (aug 2017), 1–11.